Imperial College London

DØ Computing Status and Budget Requirements

Gavin Davies
Imperial College London

IFC, October 2005

Outline

Imperial College London

- DØ Computing Model
  - Evolution of 'long' established plan
  - Evolution of associated planning tools
- Operational status / Run II Computing Review
  - Globally – continue to do well
  - Strong praise - Issues raised those we are aware of / working on
- Highlights from the last year
  - SAM-Grid & reprocessing of Run II data
    - 10^9 events reprocessed on the grid – largest HEP grid effort
- Budgetary /planning issues
- Conclusions

Apologies - Reminder of Data Flow

Imperial College London

- Data acquisition (raw data in evpack format)
  - Currently limited to 50 Hz Level-3 accept rate
  - Request increase to 100 Hz, as planned for Run IIb – see later
- Reconstruction (tmb/DST in evpack format)
  - Additional information in tmb → tmb^{++} (DST format stopped)
  - Sufficient for 'complex' corrections, inc track fitting
- Fixing (tmb in evpack format)
  - Improvements / corrections coming after cut of production release
  - Centrally performed
- Skimming (tmb in evpack format)
  - Centralised event streaming based on reconstructed physics objects
  - Selection procedures regularly improved
- Analysis (out: root histogram)
  - Common root-based Analysis Format (CAF) introduced in last year
  - tmb format remains

Computing Model - I

Imperial College London

- Started with distributed computing with evolution to automated use of common tools/solutions on the grid (SAM-Grid) for all tasks enabling physics analysis
  - Scalable
  - Not alone
- 1997 – Original Plan
  - All Monte Carlo to be produced off-site
  - SAM to be used for all data handling, provides a 'data-grid'
- Now: Monte Carlo and data reprocessing with SAM-Grid
- Next: Other production tasks e.g. fixing & finally grid-user analysis
  - 'Local' off-site analysis well established (as have SAM)
- Evolution of associated planning tools and 'virtual centre' to evaluate remote contributions – more later

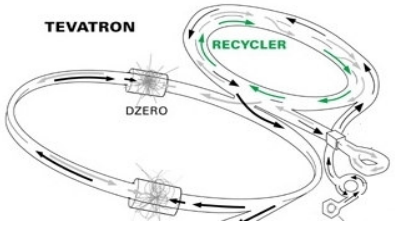
Computing Model - II

Imperial College London

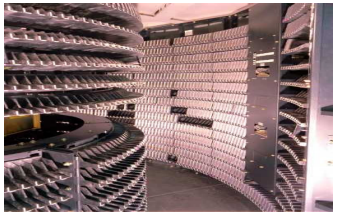
Remote Farms

Central Farms

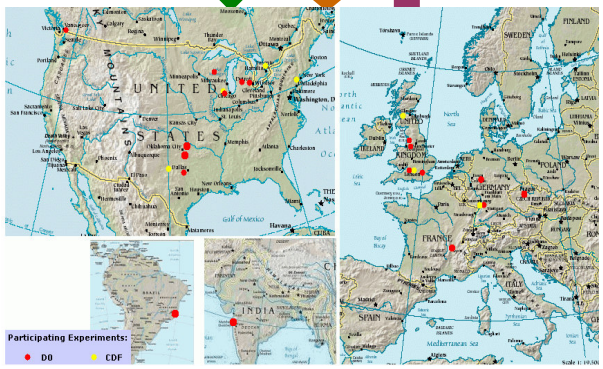
Raw Data
RECO Data
RECO MC
User Data



Data Handling Services



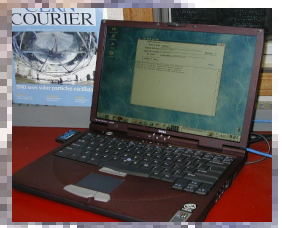
Central Storage



Remote Analysis Systems

Central Analysis Systems

Desktops



Run II Computing Review / Status

Imperial College London

- Annual review (13th-15th Sept)
  - Chaired by Jim Shank
  - CDF, DØ & computing division - ~3/4 day for DØ
  - http://cdinternal.fnal.gov/RUNIIRev/runIIMP05.asp
  - http://d0server1.fnal.gov/projects/Computing/Reviews/Sept2005/Index.html
    - including full documentation and spreadsheets
- Feedback
  - Closed session with feedback & written report on its way
  - Overall strong praise
    - Denoted with ✓
  - Points raised are areas that are already working on
    - Use review as 'ammunition'

Snapshot of Current Status

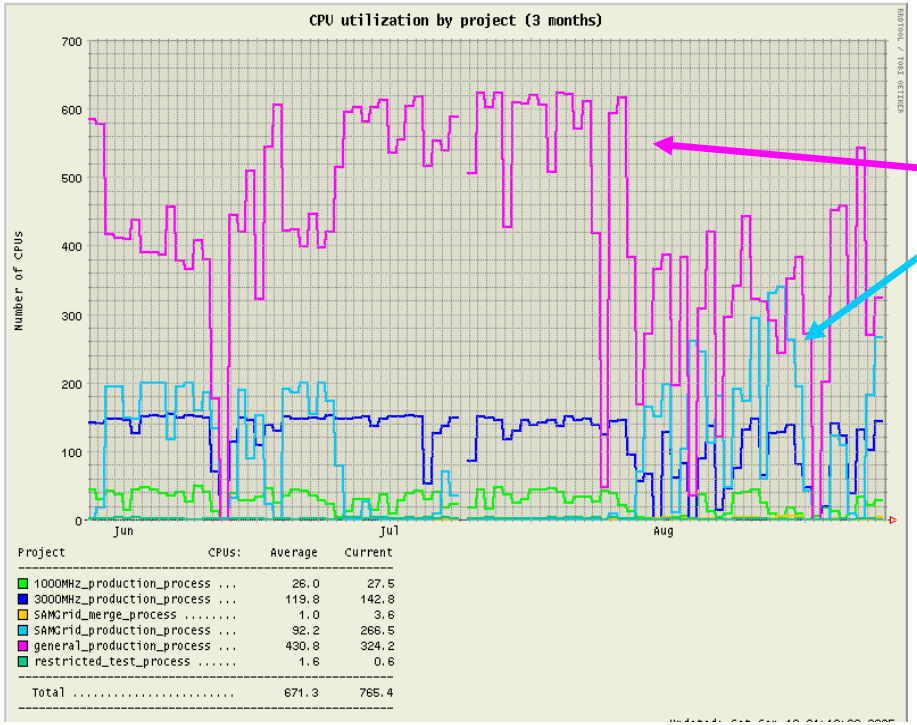
Imperial College London

- Reconstruction keeping up with data taking
  - Strong praise for d0reco speed-up, but need to keep the progress up…
- Data handling is performing well ✓✓
- Production computing: Off-site & grid based - continuing to grow & work well
  - About 80 million Monte Carlo events produced in last year
  - Run IIa data set reprocessed on the grid – 10^9 events
  - Strong praise for use of shared resources & common tools
    - Question of maintaining access to suitable resources in LHC era raised
- Common Analysis Format (CAF)
  - Simplify, accelerate analysis development & best use of resources
- Analysis cpu power has been expanded
- First $1fb^{-1}$ analyses by Moriond
  - ~$1fb^{-1}$ fixed data ready by end Nov

Globally doing well

Reconstruction

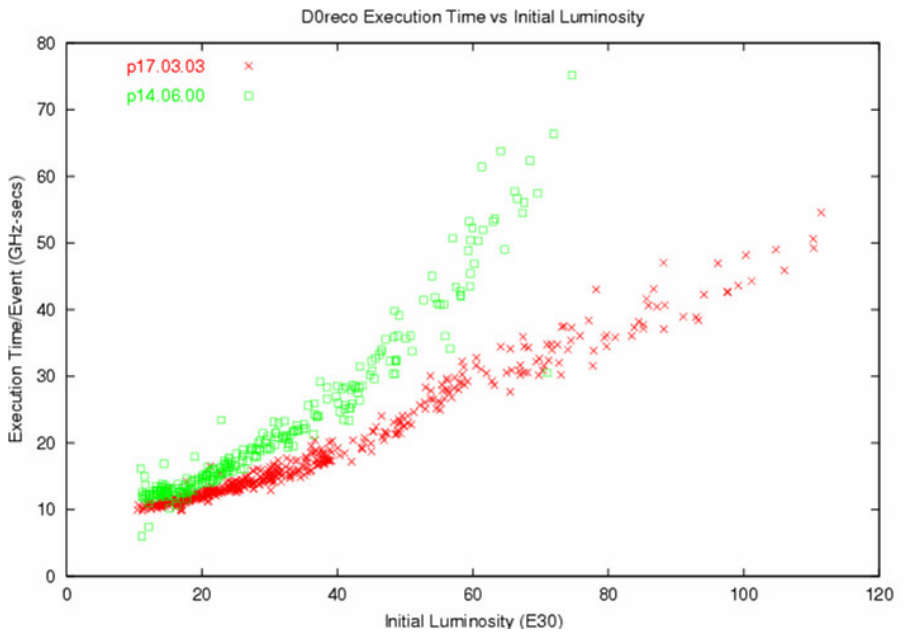
Imperial College London



Central farm

- Processing & reprocessing (SAM-Grid) with spare cycles
- Right now being used for fixing
- Evolving to shared FNAL farms

- Reco-timing
  - Significant improvement, especially at higher instantaneous luminosity

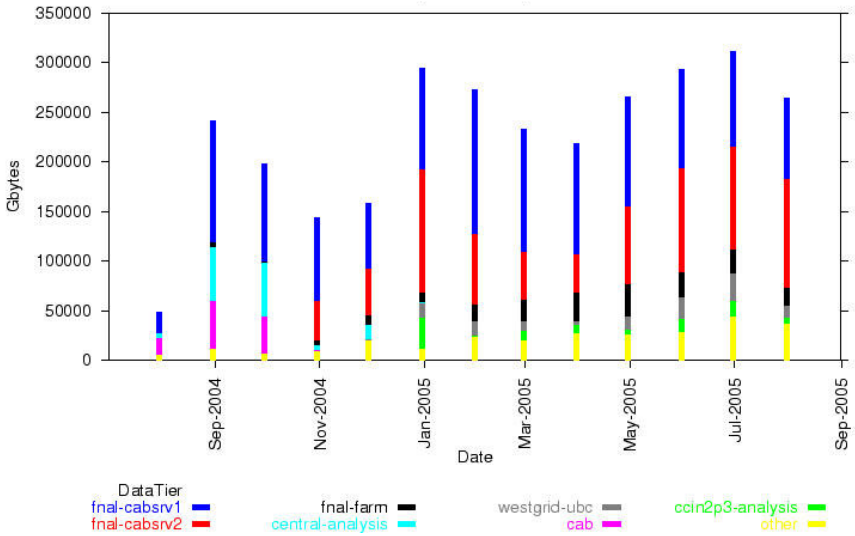


Data Handling - SAM

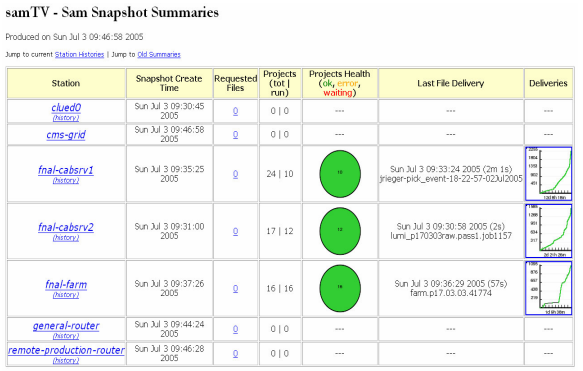
Imperial College London

- SAM continues to perform well, providing a data-grid
  - 50 SAM sites worldwide
  - Over 2.5 PB (50B events) consumed in the last year
  - Up to 300 TB moved per month

http://d0db-prd.fnal.gov/sm_local/SamAtAGlance/



  - Larger SAM cache solved tape access issues
  - Continued success of SAM shifters
    - Often remote collaborators
    - Form 1st line of defense
  - SAMTV monitors SAM & SAM stations



- SAM-Grid = SAM + JIM
  - JIM - job submission & monitoring

Monte Carlo / SAM-Grid - I

Imperial College London

- Development effort of last ~9 months has been on reprocessing, now returning to Monte Carlo – building on success of the latter
  - Consider as a single production task with common infrastructure
- Monte Carlo
  - ~80M events produced in last year, at more than 10 sites
    - More than double last year's production
  - Vast majority on shared sites
    - Often national Tier 1 sites - several "LCG"
  - SAM-Grid introduced in spring 04, becoming the default
- MC & reprocessing: Consolidation of SAM-Grid co-existence with LCG and other grids
  - ~20M events produced 'directly' on LCG via submission @ Nikhef
  - 'Full' interoperability on its way – see later

Reprocessing / SAM-Grid - II

Imperial College London

- After significant improvements to reconstruction, reprocess old data
  - P14 – winter 03/04 – from DST - 500M events, 100M off-site
  - P17 – Mid05 – from raw – 1B events – SAM-Grid default – basically all off-site
  - Massive task – largest HEP activity on the grid
    - ~3500 1GHz PIIIs for 6 months
  - Led to significant improvements to SAM-Grid
    - Collaborative effort

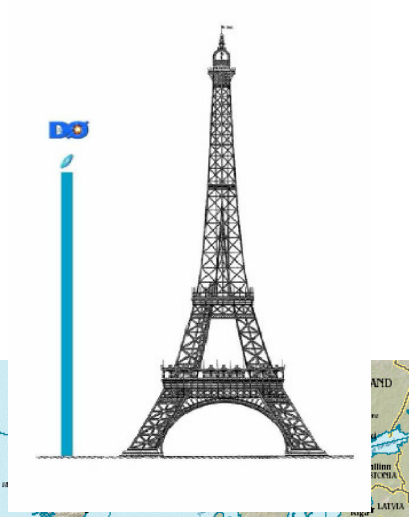
More than 10 DØ execution sites
http://samgrid.fnal.gov:8080/

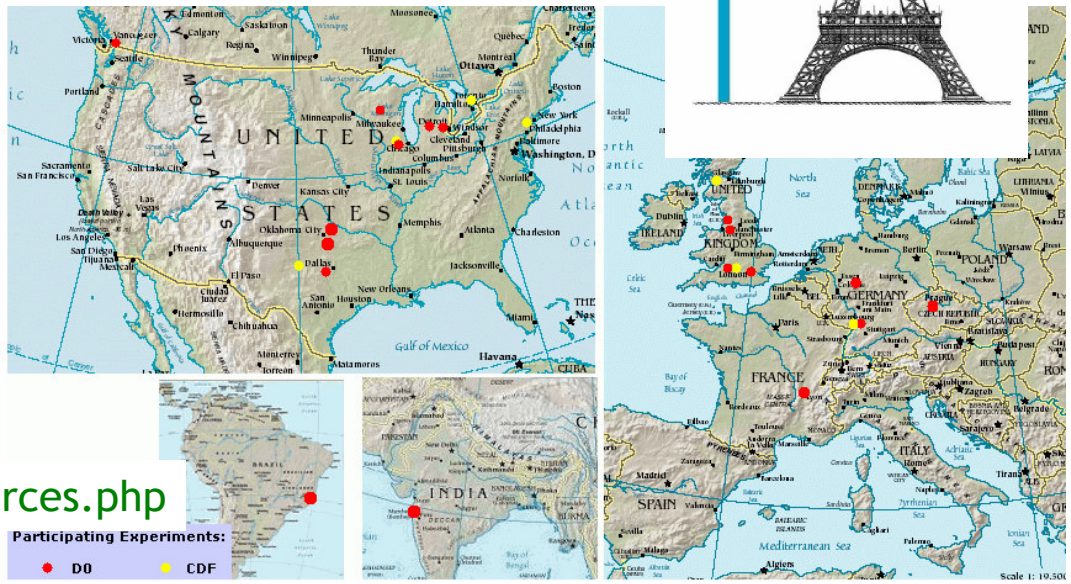
SAM Grid List of Resources

Station Name	Type	Site
caps10	SAMGRID	LTU
ccin2p3-analysis	SAMGRID	ccin2p3
ccin2p3-grid1	SAMGRID	LCG_TEST
cms-grid	SAMGRID	CMS-FNAL-WC1
d0_fzu_prague	SAMGRID	FZU_GRID
d0karlsruhe	SAMGRID	GridKa
d0ppdg-wisconsin-2	SAMGRID	Wisconsin
fnal-farm	SAMGRID	FNAL
imperial-prd	SAMGRID	IMPERIAL_PRD
luhep	SAMGRID	LUHEP
oscer	SAMGRID	OSCER
ouhep	SAMGRID	OUHEP
samgfarm	SAMGRID	SamGrid
samgfarm	SAMGRID	SamGrid
uta-rac	SAMGRID	UTA-DPCC
westgrid-ubc	SAMGRID	WestGrid

Please try clearing your browser cache and reloading the page if the information does not seem current.

http://samgrid.fnal.gov:8080/list_of_resources.php



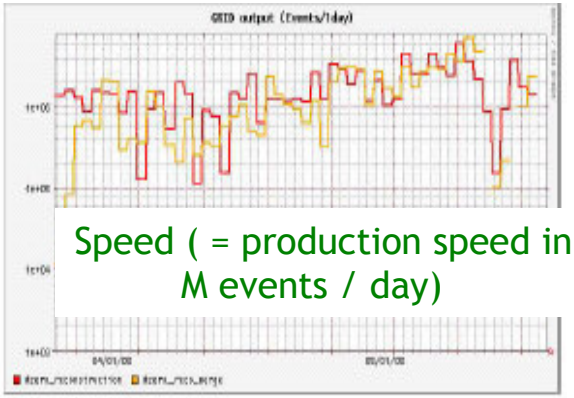


Reprocessing / SAM-Grid - III

Imperial College London

- SAM-Grid 'enables' a common environment & operation scripts as well as effective book-keeping
  - Monitor speed and efficiency
    - by site or overall

(http://samgrid.fnal.gov:8080/cgi-bin/plot_efficiency.cgi)



Speed (= production speed in M events / day)

- Started end march – ~95% done
  - In the 'cleaning-up' phase

~920M events done

http://www-d0.fnal.gov/computing/reprocessing/

P17 Reprocessing Status as of 18-Oct-2005 (all sites)

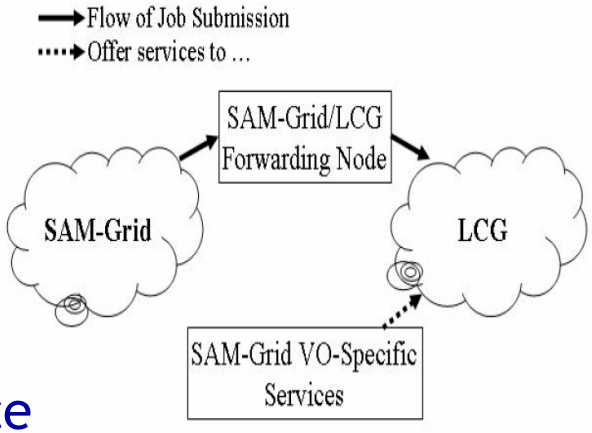
Total Raw Events	986190444	
Processed Events	922589373	
Sites	fnal, FNAL, OSCER, FZU_GRID, WestGrid, ccin2p3, GridKa, UTA-DPCC, Wisconsin, IMPERIAL_PRD, CMS-FNAL-WC1, SPRACE	

  - Comment: Tough deploying a product under evolution to new sites, as a running experiment
  - Very strongly praised at Review – further details from Daniel

SAM-Grid Interoperability

Imperial College London

- Need access to greater resources as data sets grow
- Ongoing programme on LCG and OSG interoperability
- Step 1 (co-existence) – use shared resources with SAM-Grid head-node
  - Widely done for both Reprocessing and MC
    - OSG co-existence shown for data reprocessing
- Step 2 – SAMGrid-LCG interface
  - SAM does data handling & JIM job submission
  - Basically forwarding mechanism
  - Prototype established in Fr/Germany
  - Extending to more sites & to production level



- OSG activity increasing – build on LCG experience
- Strongly praised – but limited manpower
  - Remote sites play a key role

Challenges

Imperial College London

- Issues raised by Review – things on which we were already working
- Immediate:
  - Some vulnerability through limited number of suitably qualified experts & areas where central consolidation of support → reduced overall needs.
- Increased data sets require increased resources → increased use of grid and common tools. Need effort (from both CD and expt.) for
  - Continued development of SAM-Grid
    - Automated submission of production jobs by shifters, user-grid analysis
  - Deployment team
    - Bring in new sites in manpower efficient manner
  - Full interoperability
    - Ability to access efficiently all shared resources
  - (Maintenance of production level service
    - Increased reliance on SAM-Grid places extra pressure)
- All under discussion with CD as part of FNAL taskforce

Budgetary Issues

Imperial College London

- Evolution of usual procedure
- FNAL equipment budget provides basic level of functionality
  - Databases, networking & other infrastructure, primary reconstruction, robotic storage & tape drives, disk cache & basic analysis computing, support for data access for offsite computing
- Remote Contributions
  - Monte Carlo production, reprocessing, local or collaboration wide analysis, contributions at FNAL to project disk and to CLuED0.
- Virtual Centre
  - Value: Determine the cost of the full computing system at FNAL costs, purchased in the yearly currency
  - Assign fractional value for remote contributions, using a merit based assignment of value
- Spreadsheets
  - Evolved over time, used for planning and calculating value
  - Use data rate and past experience as driving factor
    - Using metrics from SAM and system monitoring

Updates

Imperial College London

- Aggressive drive for maximum return
  - 40% speed-up of d0reco
  - Tightening of trigger and skimming criteria
  - Consolidation of data formats
    - Common Analysis Format (CAF) & suppression of DST
  - Changed to a 4-yr retirement policy
    - With 20% failure rate
  - Cost savings on infrastructure and networking
    - e.g. Replacement of d0mino, networking in FCC…..
- Looking forward
  - Stick with existing tape robot / drives (AML2 & LT0 II)
    - Will re-cycle tapes
  - Assume all major infrastructure is in place
    - Will re-use networking – budget $100k/yr
- About as lean as we can get
- Assume higher data collection rate for higher luminosity years
  - Long standing plan for Run IIb

Increased rate to tape

Imperial College London

- Experiment performing well
  - Run II average data taking eff ~83%, now pushing 90%
  - Making efficient use of data and resources
  - Many analyses published (have a complete analysis with 0.6fb^{-1} data)
- "Core" physics program saturates 50Hz rate at 1 x 10^{32}
- Maintaining 50Hz at 2 x 10^{32} → an effective loss of 1-2fb^{-1}
  - http://d0server1.fnal.gov/projects/Computing/Reviews/Sept2005/Index.html
- Feed into spreadsheets
  - e.g.

	2006	2007	2008
peak event rate	100	100	100
average event rate	34.48276	34.48276	34.48276
weekly average	50	50	50
raw data rate			
Geant MC rate	3.45	3.45	3.45

  - Combine with data sizes, MC rate, disk usage, d0reco time vs Lumi, analysis needs, skimming / fixing / reprocessing cycles…..

Cost Estimate – Sept 2005

Imperial College London

- As presented at Computing Review – under discussion now

	Purchase 2006	Purchase 2007	Purchase 2008
CPU	$449,308	$475,835	$404,720
Reconstruction	$666,665	$339,534	$156,352
File Servers/disk	$ 348,500	$328,000	$307,500
Mass Storage	$57,000	$97,500	$97,500
Infrastructure	$100,000	$100,000	$100,000
FNAL Total	$1,621,473	$1,340,869	$1,066,072

- Took guidance to be $1.5M for equipment money in 2006
- Associated operational tape costs ~ $200k /yr

- Increased rate = 1-2 fb^{-1} more physics
- Have saved where we can

- Have breakdown of table if people wish to see it.

Conclusions

Imperial College London

- DØ Computing continues to be successful
  - Significant advances this year include p17 reprocessing with SAM-Grid
  - 1fb^{-1} fixed data set ready end Nov

- DØ Computing Model continues to be successful
  - Have suitable tools, using metrics from SAM, to enable effective planning / budgeting at FNAL
  - Virtual centre calculates value of remote computing
- Continue to pursue a global vision for the best use of resources via use of automated common tools and interoperability with LCG and OSG
- Strong praise at the recent Run II review
- However: DØ computing remains effort limited – a few more skilled people could make a huge difference
- Short budgets and continued construction are also cause for concern

BACK-UP

Terms

Imperial College London

- Tevatron
  - Approx equiv challenge to LHC in “today’s” money
  - Running experiments
- SAM (Sequential Access to Metadata)
  - Well developed metadata and distributed data replication system
  - Originally developed by DØ & FNAL-CD
- JIM (Job Information and Monitoring)
  - handles job submission and monitoring (all but data handling)
  - SAM + JIM →SAM-Grid – computational grid
- Tools
  - Runjob - Handles job workflow management
  - dØtools – User interface for job submission
  - dØrte - Specification of runtime needs

Monte Carlo Statistics

Imperial College London

- e.g. Aug04-aug05

Site	Events	Size (MB)
GridKa/Wuppertal	4552800	222052
LTU	501750	24471
LU	863263	46651
OU	1618000	86907
SPRACE	3687155	191528
Tata	793800	41997
UTA	2691941	147193
Wisconsin	12778	771
CCIN2P3	32066167	1939765
FZU	7740563	385985
Lancaster	4320975	176929
Manchester	100500	6276
Nikhef/LCG	17148986	883450
TOTAL	**76098678**	**4153975**

The Good and Bad of the Grid

Imperial College London

- Only viable way to go…
- Increase in resources (cpu and potentially manpower)
  - Work with, not against, LHC
  - Still limited

BUT

- Need to conform to standards – dependence on others..
- Long term solutions must be favoured over short term idiosyncratic convenience
  - Or won't be able to maintain adequate resources.
- Must maintain production level service (papers), while increasing functionality
  - As transparent as possible to non-expert

Accumulation Estimates / Disk costs

Imperial College London

- 2006 purchases provide capacity for 2007

data samples (events)			
Current	2006	2007	2008
events collected	1.09E+09	1.09E+09	1.09E+09
total events	2.90E+09	3.99E+09	5.08E+09
Geant events	1.09E+08	1.09E+08	1.09E+08
PMCS events	1.09E+08	1.09E+08	1.09E+08
TAPE data accumulation (TB)			
Yearly storage (TB)	888	1,102	1,308
total storage (TB)	2,248	3,349	4,658
disk data accumulation (TB)			
total storage (TB)	186	186	186

Fileservers:

	2006	2007	2008
Data Volume (TB)	186	186	186
Project Volume	31	31	31
total volume	217	217	217
contingency	40%	40%	40%
years volume (# servers)	17	16	15
Cost	$ 348,500	$ 328,000	$ 307,500

- Model for cache space / disk resident samples under evolution, assume 40% contingency as in past

Primary Production

Imperial College London

- Rate increased planned as part of upgrade
- Opening up to Fermigrid

Primary Reconstruction Cost Estimate			
Year	2006	2007	2008
Average Rate	34.48275862	34.48275862	34.48275862
efficiency	80%	80%	80%
contingency	20%	20%	20%
Reco time	85	100	100
Required CPU	2092759	2462069	2462069
Existing system	902761	1704485	2025993
Nodes to purchase	208	106	49
Node Cost	$666,665	$339,534	$156,352

- Analysis cpu: Calculated in same way, using observation that weekly analysis is ~ total data set collected

Virtual Centre & Tape costs

Imperial College London

- Reflects full value of doing all DØ computing in 1yr
  - Uses current yr $ - legacy system worth replacement cost
  - Refinements continue (infrastructure , fixed value for mass storage)

	Value 2005	Value 2006	Value 2007	Value 2008	Value 2009
FNAL Based CPU	$2,192,370	$2,073,155	$2,285,221	$2,752,547	$2,429,034
File Servers/disk	$369,000	$758,500	$984,000	$1,209,500	$1,271,000
Mass Storage	$800,000	$800,000	$800,000	$800,000	$800,000
Reprocessing	$4,013,039	$4,187,340	$4,208,316	$6,526,760	$5,438,967
MC	$436,484	$219,458	$234,089	$187,271	$149,817
Center Total	$7,810,893	$8,038,454	$8,511,625	$11,476,078	$10,088,818

- Tape Costs – part of operating budget

		2005	2006	2007	2008
Data Volume		629	888	1,102	1,308
# to retire		0	0	0	0
Tape Cost		$ 157,250	$ 177,680	$ 192,815	$ 229,005

- Staying with LT0 II driven by large saving in equipment cost
  - Savings due to recycling tapes not shown

Infrastructure Costs

Imperial College London

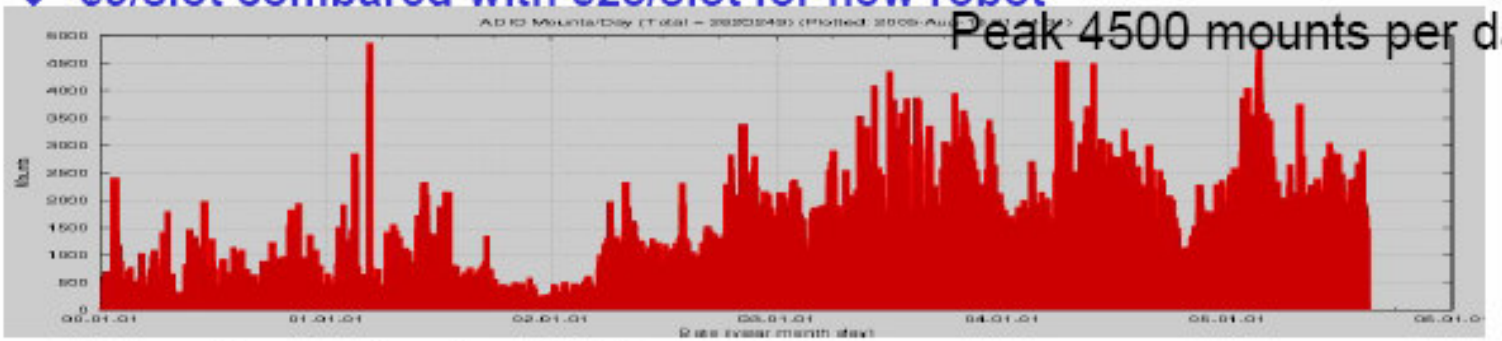
- FY-2005 Replace aging components
  - Retired Domino-> replace with login pool
    - Budgeted $100K, cost $30K
  - Home areas SGI D02ka -> Network appliance
    - Budgeted $100K, cost $63K
  - Purchased new db machines for luminosity db, added disk
    - Budgeted $100K, cost $70K
  - Networking-buying dual core worker nodes and running cables from FCC1 to FCC2
    - Budgeted $225K->cost $130K
  - We worked aggressively to bring the costs down.
  - Budget $100K per year—reuse networking

Amber Boehnlein, FNAL

Mass Storage

Imperial College London

- DO uses STK powerhorn silos and an ADIC AML/2
- Have 16 9940b drives, 14 (+4) LTOII drives
- 1/3 of files consumed for analysis can be transferred from tape (compare 2/3 at peak in 2004)
  - Activated the second arm in the AML/2
- In 2005 "traded" 9940b drives for 4 LTOII drives, have 3500 STK Slots. Use for D0 raw data, accommodate CDF need
- Plan to remain with AML2 and LTO II, will have to activate the third quadrotower
  - Currently sharing the AML/2 with SDSS
  - $9/slot compared with $25/slot for new robot



Mounts per day in AML/2, past 5 years.

Amber Boehnlein, FNAL